Visual and Auditory Characteristics of Talkers in Multimodal Integration

A Senior Honors Thesis

Presented in partial fulfillment of the requirements for graduation with distinction in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

by

Kyle Shepard

The Ohio State University

May 2009

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

In perceiving speech, there are three different elements of the interaction that can affect how the signal is interpreted: the talker, the signal (both the visual and auditory) and the listener. Each of these elements inherently contains substantial variability, which will, in turn, affect the audio-visual speech percept. Since the work of McGurk in the 1960s, which showed that speech perception is a multimodal process that incorporates both auditory and visual cues, there have been numerous investigations on the impact of these elements on multimodal integration of speech. The impact of talker characteristics on audio-visual integration has received the least amount of attention to date. A recent study by Andrews (2007) provided an initial look at talker characteristics. In her study, audiovisual integration produced by 14 talkers was examined, and substantial differences across talkers were found in both auditory and audiovisual intelligibility. However, talker characteristics that promoted audiovisual integration were not specifically identified. The present study began to address this question by analyzing audiovisual integration performance using two types of reduced-information speech syllables produced by five talkers. In one reduction, fine-structure information was replaced with band-limited noise but the temporal envelope was retained, and in the other, the syllables were reduced to a set of three sine waves that followed the formant structure of the syllable (sine-wave speech). Syllables were presented under audio-visual conditions to 10 listeners. Results indicated substantial across-talker differences, with the pattern of talker differences not affected by the type of reduction of the auditory signal. Analysis of confusion matrices provided directions for further analysis of specific auditory and visual speech tokens.

Acknowledgments

I would like to thank my advisor, Dr. Janet M. Weisenberger, for her constant support and guidance she has given me throughout this process and for the past two years. She has been vital in my development as a student, a researcher, a professional and as a person. I would like to thank Michelle Hungerford for all of the assistance and direction she provided us in the lab. I would also like to thank my mentor and boss Dr. Sharla Wells-Di Gregorio. She is the reason I became interested and involved in research. It is because of her that I know I want to be both a clinician and a researcher in the future. Furthermore, I would like to thank all of my subjects for their patience and dedication while assisting me with my thesis.

This project was supported by an ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

Speech perception is generally thought to occur primarily through our sense of hearing. However, an individual's ability to understand speech is a multimodal process that incorporates both auditory and visual cues. When an auditory signal is impoverished, whether it is due to a hearing loss or a noisy situation, people use visual inputs to compensate for auditory loss. In addition, even when the auditory signal is highly intelligible, it has been demonstrated by that visual information will still be used and affect what is construed (McGurk & MacDonald, 1976).

The McGurk and MacDonald study dubbed auditory syllables onto a video with a talker visually saying a conflicting visual stimulus. The result was listeners perceiving a fusion or combination of the two syllables to form what is now known as the McGurk Effect. Auditory stimuli were paired with conflicting visual stimuli to determine how participants would integrate different audio and visual inputs, as well as to explore whether one modality would dominate the other in the perception of speech. Different presentations consisted of pairing the auditory /ba/ with the visual /ga/ (b/g), the auditory /pa/ with the visual /ka/ (p/k), the auditory /ma/ with the visual /da/ (m/d), and the auditory /va/ with the visual /da/ (v/d) (Grant and Seitz, 1998). When the visual syllable /ga/ was simultaneously presented with the auditory syllable /ba/, the result was listeners perceiving a fusion of the two syllables, /da/ or /ða/. Essentially, the glottal phoneme /g/ was fused with the bilabial phoneme /b/ to form the alveolar phoneme /d/ or /ð/, an intermediate place of articulation. When the auditory and visual (g/b) syllables were swapped, listeners grouped the two inputs to perceive a combination response of /bga/. Because the visual stimulus /ba/ is very salient, fusion with the auditory /ga/ did not occur. McGurk and MacDonald

were the first to demonstrate that auditory-visual integration is a natural component of speech perception that is inherently utilized in any situation, even when only one modality is required.

When perceiving speech there are three different components of the interaction that can affect how the signal is interpreted: the talker, the signal (both the visual and the auditory) and the listener. Each of these three components can have huge variability, which will, in turn, alter the audio-visual speech perception. Since the McGurk study, there have been numerous investigations addressing these components and the effect each has on multimodal integration.

Auditory Cues for Speech Perception

A great deal of information can be extracted from the auditory signal to help a listener perceive speech, including cues for place, manner, and voicing in both the temporal and spectral structure of the waveform. Place of articulation is cued by formant transitions and refers to the location in the oral cavity where the articulation of speech occurs. The different places of articulation include bilabials (with the lips), labiodentals (with the lower lips and upper front teeth), interdentals (with the tongue between the teeth), alveolars (with the tip of the tongue and the alveolar ridge), palatal-alveolars (with the blade of the tongue and the alveolar ridge), palatals (with the tongue and the hard palate), and velars (with the tongue and the soft palate). Manner of articulation is cued through formant intensity and formant frequency changes and describes how the articulators make contact with each other as sound flows through the oral cavity in the production of speech. Different manners of articulation include stops, fricatives, affricates, liquids, and glides. Voicing is cued by voice onset time (VOT), the length of time that passes between the release of a consonant and the vibration of the vocal folds, and refers to the

state of the vocal folds as a sound is being produced. Voiced sounds are sounds produced while the vocal folds are vibrating and voiceless sounds are sounds produced while the vocal folds are at rest (Behrman, 2007).

A number of studies have shown that a substantial amount of information can be removed from the speech signal without significantly reducing intelligibility. In a study conducted by Shannon *et al*. (1995), it was shown that speech can still be highly intelligible with greatly reduced spectral information. Shannon focused on degrading selected aspects of the speech waveform in a manner similar to that employed in cochlear implant processors. Temporal and amplitude cues were preserved in each spectral band while the spectral detail within each band was replaced with band-limited noise (Shannon *et al*., 1995). Results showed that although speech identification improved as the number of noise bands modulated by the speech temporal envelope increased, speech was still highly recognizable with only three bands of modulated noise. This study was able to show how a non-degraded auditory signal is highly redundant, in the sense that it contains much more information than needed to recognize speech. While all the information contained in a speech waveform is not useless, it also is not essential to obtain high speech recognition performance.

Shannon and colleagues expanded on this research in 1998 to identify which parameters of sound, when reduced, are most important for recognizing consonants, vowels, and sentences. It was found that consonants were much less affected by spectral distributions of the envelope cues when compared to vowels (Shannon *et al*., 1998). Even though Shannon and his colleagues found that consonant phonemes can be easily perceived when the sound is highly degraded, speech recognition is still very difficult because of the effect on vowel recognition.

Remez and his colleagues performed a similar study to Shannon's investigation of redundancy in the speech signal in 1981. However, the acoustic waveforms were degraded differently. In this study, Remez *et al*. reduced the acoustic signal into three time-varying sinusoidal patterns that followed changing formant center frequencies of a naturally produced utterance, known as sine wave speech. Three independent groups of listeners were presented with different degrees of information about the stimuli they would hear. The first group was asked to give their spontaneous impressions of the sounds they would be hearing without being told anything about the nature of the sound. The second group was told they would hear a computer generated sentence and were asked to transcribe the utterances to the best of their ability. The final group was given the sentence they would hear, "Where were you a year ago?" and asked to evaluate the speech quality of this sine wave speech while also judging how intelligible the sentence was (Remez *et al.*, 1981). Results showed a dependence on the amount of information listeners received about the stimulus. When specific knowledge about the stimulus being presented was provided, even with a considerable amount of information missing from the speech signal, the listeners were able to identify and describe the utterances. However, when individuals were not informed that the stimulus was speech, they did not automatically perceive sinusoidal replicas of natural speech as linguistic entities. Even though results showed that reducing the signal in this manner may not be ideal for spontaneous perception, Remez demonstrated that speech perception can endure some absence of acoustic and formant cues, as long as the natural speech pattern is preserved.

Visual Cues for Speech Perception

While auditory cues for articulatory features such as place, manner and voicing are critical in perceiving speech, there are also beneficial cues contained in the visual input. In contrast to auditory cues, which provide several articulatory components, visual cues primarily provide information regarding place of articulation. A significant amount of information can be pulled from these visual cues through the display of movement in the talker's eyes, mouth and head (Munhall *et al.*, 2004).

Because cues to manner and voicing are less evident in the visual signal, the listener encounters problems when speech sounds are not distinguishable from one another based on the visual stimulus alone. Some phonemes can be easily distinguished solely based on visual differences, whereas other groups of phonemes cannot be discerned by vision alone because of the similarity in the visual movements in their production. For example, the phonemes /p/, /b/, and /m/ all differ auditorily but normally cannot be distinguished visually because they are all produced as bilabial consonants. These similar phonemes are called visemes or groups of sounds possessing the same visual features (Fisher, 1968). Viseme groups allow speechreaders to distinguish between groups of sounds but not between individual phonemes within the group (Jackson, 1988). This can present a problem to the listener when relying on visual cues alone.

Although visual cues can be very useful for speechreaders, they are also very dependent on talker differences. Viseme groups can vary greatly based on the way a given talker produces sound. Due to the vast diversity of talkers, there are no universal visemes. Jackson (1988) found that talkers who were easier to understand produced more viseme groups when compared to more difficult talkers, who produced fewer viseme groups. Nitchie (as cited in Jackson, 1988)

used the term “homophenous” for speech sounds that are visually indistinguishable because they share the same place of articulation. “Homophenous” only applies to consonants because every English vowel is produced with a distinct shape. Homophenous sounds can create words whose production looks identical despite sounding and being spelled differently, such as bed and pet (Schow & Nerbonne, 2007)

Auditory-Visual Integration Theories

Other research has focused on how the auditory and the visual signal are used together in the perception of speech. “Audio-visual integration” refers to the processes utilized by receivers to combine information extracted from auditory and visual sources (Grant, 2002). A variety of models have been developed to describe this process of integration across modalities for optimal speech perception. Grant (2002) discusses two of these models, the Fuzzy Logic Model of Perception and the Prelabeling Model of Integration, attempting to determine their success in predicting integration.

The Fuzzy Logic Model of Perception (FLMP) implies that incoming auditory, visual and audio-visual information is independently evaluated by listeners by comparing summary descriptions of the incoming signal with known descriptions in the memory to determine the degree to which the cues from a given source match alternative responses. From these alternatives, a decision is made based on the amount of support contained in the memory for each response possibility (Massaro, 1987 as cited in Grant, 2002). Massaro (1987) states that the multiplicative integration rule used in the FLMP is an optimal decision rule used to minimize differences between obtained and predicted scores, and is therefore considered more of a fit to

obtained bimodal scores rather than a prediction of optimal bimodal speech performance. Grant (2002) recognized two consistencies demonstrated in the FLMP relating to this concept: first, it seeks to apply multiplicative integration to unimodal data (i.e., probabilities of responding *y* given *x*) to obtain a bimodal prediction, and second, human receivers often do better at recognizing consonants than the FLMP predicts. The FLMP is presumed to be a model that predicts optimal integration (Massaro, 1987; Massaro and Cohen, 2000, as cited in Grant, 2002). Therefore, poor multimodal results should be attributed to poor unimodal inputs and not to poor integration abilities.

Contrasting to the FLMP, the prelabeling model of integration (PRE) does not seek to optimally fit observed auditory-visual data, but instead seeks to "label" incoming bimodal stimuli based on an optimal combination of mutual information acquired from separate fits to auditory-only and visual-only performance (Braida, 1991, as cited in Grant, 2002). This model first obtains an estimate of unimodal information and then predicts how an unbiased receiver with no interference across modalities might perform given the particular unimodal information available by using an optimum combination rule. Unlike the FLMP, which assumes optimal integration, the PRE allows for the possibility that integration ability may be suboptimal. The PRE model was determined to be a better fit in predicting integration because of the model's ability to account for individual differences seen in the speech perception of hearing-impaired individuals who participated in the Grant and Seitz (1998) study. It is important to realize that just because the PRE provided a better fit to the data, this does not mean that it is a more valid assessment of integration efficiency than the FLMP. Rather, the two models focus on different aspects of the integration process. The PRE further serves as an example of an "early" integration theory, in that the visual and auditory inputs are combined prior to the decision stage,

whereas the FLMP serves as an example of a "late" integration model, in which unimodal decisions are combined to yield a final decision.

Grant (2002) emphasizes that it is also important to realize that audio-visual integration is a process that combines information pulled from auditory and visual sources and differs considerably from a listener's actual ability to extract auditory and visual cues and higher-order processing of the information received by the two senses (Massaro, 1998 as cited in Grant, 2002). While research has continually shown that audio-visual integration abilities of people can be very resistant to different conditions of signal distortion, there are still many individual differences in auditory-visual speech perception performance (Grant, 2002).

Grant and Seitz (1998) documented these individual differences among listeners in auditory-visual integration. Their study examined the variability of integration abilities across individuals through comparing the previously described audio-visual integration measures (PRE and FLMP). Their results were similar to those of other studies, in that speech was still recognizable even with a significantly degraded auditory signal when it is paired with a visual stimulus. More importantly, it was found that two listeners, who are equally good/bad in either auditory-only or visual-only conditions, vary considerably when it comes to audio-visual integration efficiency. Factors such as hearing loss, visual acuity, vocabulary, and language competence were controlled for in the administration of various auditory-visual speech recognition tasks. The results strongly support that audio-visual integration is a unique, independent process with huge individual differences.

Recent Auditory-Visual Integration Research

Several recent studies in our laboratory have investigated the different components of audio-visual perception of reduced-information speech stimuli (i.e. the listener, the signal and the talker).

Listener Characteristics

Feleppelle (2008) examined degraded auditory signals degraded in a manner similar to that of Shannon *et al.* (1995) in an audio-visual integration task. Speech syllables were degraded using 2, 4, 6, and 8 bandpass filter channels. Results showed that that as the auditory signal became further degraded, performance in both the auditory-only and the audio-visual conditions decreased. However, the amount of integration remained relatively constant. These results suggest that although audio-visual integration can still be accomplished with a degraded auditory signal, the auditory signal does play a role in the intelligibility of speech. As seen in Grant and Seitz (1998), Feleppelle also observed individual differences among listeners in audio-visual integration of a degraded auditory signal.

Auditory and Visual Speech Stimuli Characteristics

Huffman (2007) altered characteristics of the auditory signal and observed the impact on auditory-visual integration. By isolating and systematically removing progressively greater amounts of information from the auditory signal, she explored whether acoustic redundancy or ambiguity better facilitated optimal auditory-visual integration (Huffman, 2007). The auditory

stimuli used in this study were also similar to the stimuli degraded by Shannon *et al*. (1995). Huffman found that some of the place, manner, and voicing cues may be lost due to the noise fine structure of these reduced speech signals. Results of this study supported the fact that listeners perform better when more auditory information is available; however, removing information from the auditory stimulus did not affect the degree of integration benefit. Substantial across-talker differences were also observed in auditory intelligibility in the 2-channel condition. Additionally, the degree of audiovisual integration produced by different talkers was unrelated to auditory intelligibility (Huffman, 2007).

Dietrich (2008) did a preliminary evaluation of the acoustic characteristics that were most important for identifying 2-channel and 4-channel stimuli similar to those used by Huffman. Her analysis indicated that clear F2 formant transitions from the initial consonant to the medial vowel were a primary determinant of intelligibility.

Similar to Huffman (2007), Tamosiunas (2007) investigated how a reduced auditory signal affects audiovisual integration. However, he degraded the signal in a manner similar to that of Remez *et al.* (1981). He was interested in whether reducing the redundancy in the auditory signal changes the audiovisual integration process in either qualitative or quantitative ways (Tamosiunas, 2007). In contrast to Remez's investigation of redundancy in the speech signal that looked at intelligibility in sentences, Tamosiunas looked at integration performance for isolated CVC syllables using sine wave speech. Results showed that for isolated syllables, sine wave reduction of speech effectively reduces the available acoustic information found in the signal. This finding was in contrast to Remez's study, which found sine wave speech to be highly intelligible when identifying sentences. Furthermore, these results contradict Huffman's findings, in that his reduced auditory stimuli (sine wave speech) actually impeded integration instead of

facilitating it. Tamosiunas also suggested that there might not be enough information contained in individual sine wave speech syllables to facilitate optimal audiovisual integration (Tamosiunas, 2007).

Building from Huffman and Tamosiunas's studies, Hiss (2008) performed a within-subjects comparison of intelligibility for both types of auditory reduction previously described (2-channel and sine wave speech). Hiss was concerned that the varying results between the two studies could be due to the fact that they used different groups of subjects and not because of the type of stimulus reduction. Results showed that participants performed far better with 2-channel filtered speech than sine wave speech. However, subjects showed more audiovisual integration with sine wave speech, suggesting that a more highly degraded auditory stimulus promotes greater integration (Hiss, 2008).

Talker Characteristics

Although there have been several studies that examine some of these key components in perceiving speech, aspects relating to characteristics of talkers in audio-visual integration are still under-researched. For example, the visual and auditory characteristics that make a talker either good or bad are still unclear. Does a perfect audio talker produce the best audiovisual integration? Does a poor audio talker provide the best integration? Are there certain visual cues that promote the best multimodal integration? An initial look at these questions was performed by Andrews (2007). She studied the amount of audiovisual integration produced by 14 talkers to examine talker differences using 2-channel filtered speech. Substantial differences across talkers were found in both auditory and audiovisual intelligibility. Similarly, Anderson (2007) examined talker differences in audiovisual integration produced by the same 14 talkers in Andrews (2007)

study. Anderson, however, used sine wave speech reduction. Substantial differences across talkers were also found in both auditory and audiovisual intelligibility. However, the characteristics that contributed to across-talker differences were not specifically examined or discussed in either study. For example, why do two talkers, who produce very similar scores for auditory-only and visual-only intelligibility, produce very different audiovisual performance? In addition, neither study addressed whether the pattern of talker differences could be attributed to the type of auditory reduction. For example, would a highly intelligible talker in 2-channel filtered speech still be a highly intelligible talker in sine wave speech or would across talker differences change according to the type of signal reduction?

The present study began to address those issues. Focus was put on the type of auditory reduction and whether it affects patterns of talker differences. Other possible reasons for across-talker differences were also examined and discussed. Audiovisual stimuli produced by five talkers were presented to a group of ten normal hearing listeners for identification. Auditory stimuli were similar to the stimuli degraded by both Shannon *et al*. (1995) and Remez *et al*. (1981) (i.e., 2-channel filtered speech and sine wave speech). Statistical analysis allowed us to analyze talker differences, speech type differences and whether speech type contributed to talker differences. Confusion matrices from listener responses allowed us to evaluate differences in specific perceptual confusions across talkers and assess whether the specific form of auditory information reduction is an important factor.

A better understanding of how talker differences impact audiovisual integration should offer some guidance for development of aural rehabilitation programs for hearing impaired perso

Chapter 2: Methods

Participants

Participants in this study included ten listeners. They included seven females and three males, with ages ranging from 20 to 23. All ten listeners reported having normal hearing and normal or corrected vision. They each received $70.00 for their participation in this study. Previous digital video recordings of five talkers provided the stimuli used in the present study. The talkers used consisted of three females and two males. These talkers have been used in past studies focusing on audiovisual integration. Talkers were chosen to include two who were highly intelligible in previous research and three who were much less intelligible. All five talkers reported being Native English speakers. They were not compensated for their participation.

Interfaces for Stimulus Presentation

Visual Presentation

Visual stimuli were presented via a 20" video monitor connected to a DVD player. The monitor was positioned approximately four feet away from the participant's head at eye level.

Auditory Presentation

Auditory stimuli were presented via TDH-39 headphones at approximately 75 dB SPL.

Stimuli Selection

A set of eight CVC (consonant-vowel-consonant) syllables were presented as stimuli for this study. These syllables included:

1. bat
2. cat
3. gat
4. mat
5. pat
6. sat
7. tat
8. zat

The four following dual-syllable (dubbed) stimuli were also used to elicit McGurk-like responses. The first column represents the visual stimulus and the second column represents the auditory stimulus:

1. bat-gat
2. gat-bat
3. cat-pat
4. pat-cat

These syllables were chosen to satisfy the following conditions:

1. Pairs of the stimuli were minimal pairs, differing only in initial consonant.
2. All stimuli were accompanied by the vowel /ae/, which does not exhibit lip rounding or lip extension.
3. Multiple stimuli were used in each category of articulation, consisting of: place (bilabial, alveolar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced).

4. All stimuli were presented without a carrier phrase.
5. Stimuli were known to elicit McGurk-like responses when appropriately chosen pairs of syllables were combined and presented properly.

Stimulus Degrading and Editing

Each talker was recorded with a digital video camera while they produced a set of eight monosyllabic words five times each. Their voices were recorded through a microphone that was directly connected to a computer. This allowed the stimuli to be saved as files in .wav format. The auditory stimuli were degraded in two different ways, one being a 2-channel filtered stimulus and the other being 3 formant sine wave speech.

Audio Signal Degrading

2-Channel Filtered Speech

For the 2-channel speech, the auditory files were converted into degraded auditory speech samples using a MATLAB subroutine created by Bertrand Delgutte (Smith, Oxenham & Delgutte, 2002). This program takes the input speech waveform, filters it into the desired number of bands, and swaps the fine structure of each band with that of a broadband noise, while retaining the amplitude envelope characteristics. The speech signals were first filtered into two broad spectral bands, providing equal spacing in basilar membrane distance. The cutoff frequencies for the two spectral bands were 80 Hz to 1877 Hz and 1877 Hz to 19.2 kHz. After filtering, the fine structure swap was performed. The waveform containing speech fine structure and noise envelope was discarded. The remaining auditory stimuli are essentially similar to those used by Shannon et al. (1998), where the stimuli are reduced to a waveform

consisting of noise fine structure that is modulated by the temporal envelope of the original speech stimulus.

3 Formant Sine Wave Speech

For 3 formant sine wave speech, the auditory files were converted into degraded auditory speech samples using Praat Version 4.4.29, with a software script created by Chris Darwin of The University of Sussex. This script produces sine wave speech by reducing auditory files to three sine waves that represent the first three formants (F1, F2, and F3) of the original signal. The gender of the talker was considered when converting the auditory files to sine wave speech. The upper formant limits used were 5500 Hz for an adult female and 5000 Hz for an adult male. Degrading auditory signals into sine wave speech is done without adding noise to the signal, unlike the 2-channel filtered speech.

Digital Video Editing

Once all of the auditory stimuli were degraded, the program Video Explosion Deluxe was used to edit both sets of stimuli (auditory and visual). Using this program, any auditory stimulus can be dubbed onto any visual stimulus. This allowed for the visual clips to be paired with degraded auditory clips from the same talkers. Furthermore, McGurk-like stimuli could be created by taking a degraded auditory "bat" from a talker and dubbing it on a visual "gat" produced from the same talker. Randomized lists of sixty stimuli were made where degraded auditory clips were randomly paired with appropriate visual clips. The stimulus clips could then be burned to a DVD using the software program Sonic MY DVD. Four DVDs were made for each of the five talkers in each condition (2-channel and sine wave speech). A total of forty DVDs were used in the present study.

Procedure

Testing Setup

Testing for this study was done in a laboratory in The Ohio State University's Speech and Hearing Science department. Each participant was tested individually in a sound-attenuating booth with the door closed. The participants wore TDH-39 headphones and watched a television monitor located outside of the booth approximately four feet away. An intercom system was located inside the booth, allowing the examiner to listen to and record the participant's responses. Testing lasted approximately seven hours for each participant, and was broken up into one or two hour sessions.

Testing Presentation

All participants were given a set of instructions to read. These instructions explained that each stimulus being presented would end in "at" and they would use a closed-set response list to choose from: bat, pat, mat, gat, cat, zat, tat, sat, dat, nat, bdat, pcat, bgat, and ptat. The instructions stressed the importance of verbally responding to what they perceived on the video monitor and/or through the headphones and that there would be approximately three seconds between stimuli. Forty DVDs were randomly presented to each subject, each lasting approximately seven minutes. Each of the five talkers had four DVDs created for each auditory condition (2-channel and sine wave speech). Every DVD contained sixty randomized trials. Half of trials were congruent stimuli (same auditory and visual stimulus) while the other half were discrepant stimuli (different auditory and visual stimulus) created to elicit McGurk-like responses. Every trial presented was an auditory-visual stimulus; no auditory-only or visual-only stimuli were presented.

Chapter 3: Results and Discussion

Two types of stimuli were analyzed in this study. First, percent correct performance was assessed for single-syllable/congruent stimuli (same visual and auditory stimulus). Second, percent response was measured for dual-syllable or discrepant stimuli created to elicit McGurk-like responses (different visual and auditory stimulus). There are no "correct" responses for these discrepant stimuli, but the responses are categorized into one of three groups: "auditory," where the response is identical to the auditory stimulus used in the discrepant pairing; "visual," where the response is identical to the visual stimulus used in the discrepant pairing; or "other," where the response matches neither the auditory nor visual stimulus used in the discrepant pairing. The "other" responses are then analyzed for the occurrence of McGurk-like responses where listeners integrate the differing visual and auditory stimuli to produce a "fusion" or "combination" response.

Congruent Stimuli

As previously stated, congruent stimuli were analyzed for percentage correct performance because there was only one correct answer for these stimuli. Figure 1 shows the overall percent correct performance averaged across listeners for each of the five talkers for both the 2-channel and sine wave speech conditions. There are several things to note from this figure. First, it is obvious that listeners performed substantially better in the 2-channel filtered speech condition, for all talkers, which supports previous findings (e.g., Hiss, 2008). A 2-factor repeated measures ANOVA performed on arcsine-transformed data confirmed a significant main effect of speech type, $F(1,9) = 30.23$, $p<.001$. A significant main effect of talker was also found, $F(4,36) = 31.42$, $p<.001$. Talkers who were more intelligible in previous studies were similarly more intelligible in the present study. Likewise, talkers who were less intelligible in previous studies

were also less intelligible in this study. Means comparisons indicated significant differences between talkers DA & LG, DA & EA, KS & LG and JK & LG. Though substantial talker differences were observed for both types of speech, no significant interaction effect was found between talker and speech type. Therefore, the pattern of talker differences was the same for both 2-channel and sine wave speech. This suggests that the characteristics that make an individual a highly intelligible talker remain intact for very different forms of degradation of the acoustic stimulus, answering our initial question about whether reduction type had an effect on talker difference patterns. Although an interaction effect would have provided insight on where further examination might best reveal the factors underlying talker differences, the lack of a significant interaction indicates that some talker differences are not specific to the way in which the auditory signal is reduced.

Tables 1 and 2 contain confusion matrices showing percentage response for each of the eight congruent stimuli, averaged across talkers and listeners. Table 1 represents the 2-channel filtered speech condition while Table 2 represents the sine wave speech condition. A general finding in both matrices is that listeners' responses usually fell into the same viseme category as the correct answer, suggesting that visual information was very salient. 2-channel speech confusions tended to share manner of voicing characteristics as well as place of articulation, indicating that this type of auditory information was available. For velar consonants, some confusions fell into the alveolar category most likely because of the lower visibility of the velar place of articulation. Also, for both types of speech the syllable "tat" was least well identified. A closer examination of the acoustic characteristics of this syllable might be useful. Results show substantially poorer performance for sine wave speech stimuli that have greater high-frequency components (e.g., voiceless stops and fricatives: pat, cat, tat and sat). This suggests that sine

wave speech reduction differentially affects high speech frequencies. Confusion matrices for individual talkers for each type of speech are found in the Appendix.

Discrepant Stimuli

Figures 2 and 3 show percent response for dual-syllable testing stimuli in the 2-channel (Figure 2) and sine wave speech (Figure 3) conditions for each of the 5 talkers. The figures are titled "Modality Reliance" because visual responses occurred significantly more often than the other response types in both auditory reduction conditions, indicating a heavy reliance on the visual modality. Not surprisingly, results demonstrate that when the auditory signal is impoverished, listeners rely heavily on visual cues. It was interesting to observe the talker differences displayed in these figures, especially when it came to the percentage of "other" responses. Some talkers, such as DA in Figures 2 and 3 and EA in Figure 3, produced more "other" responses, which once again were usually McGurk-like responses. This finding indicates the variability across talkers in audiovisual integration.

Figures 4 and 5 show percentage McGurk-like or "other" responses categorized into one of three groups: fusion, where the response was a "fusion" of the differing auditory and visual stimuli to produce an intermediate place of articulation, such as /da/ for the visual stimulus /ga/ and auditory stimulus /ba/; combination, where the response was a combination of the differing auditory and visual stimuli, such as /bga/ for the visual stimulus /ba/ and auditory stimulus /ga/; or neither, where the response was not the presented visual or auditory stimulus, nor was it a fusion or combination response. Fusion responses were the most common response across all 5 talkers in both auditory conditions. It is important to note that the number of total "other"

responses (fusion, combination and neither) differed so greatly for each talker, that percentage response should not be compared across talkers in these figures. In other words, some talkers produced few "other" responses, whereas, other talkers produced a larger number. The use of percentages, which is helpful for comparison purposes, obscures this fact. Results indicate that listeners are much more likely to provide a fusion response than a combination response, which is consistent with past research. This is in part because combination responses are not naturally permissible syllables in English. In addition, it is very easy to confuse place of articulation when the syllable produced is velar or alveolar because of the much less visible position of the articulators.

Table 3 shows the number of McGurk-like responses out of total number of "other" responses for each talker and each type of speech. Hiss (2008) found more integration for the more ambiguous auditory stimuli (sine wave speech) in her study. The present study found more integration for 2-channel speech for three of the talkers and then more integration for sine wave speech for the other two talkers. This indicates that the type of auditory stimuli doesn't fully determine the amount of integration one will receive. Differences in our results could be contributed to subject differences or differences in talkers used.

Chapter 4: Summary and Conclusion

Overall, and consistent with previous research, 2-channel filtered speech was significantly more intelligible than sine wave speech, for all talkers. Talker differences were evident in all analyses. Some talkers, such as LG & EA, were more intelligible overall, also supporting past research (Andrews, 2007; Anderson, 2007). In answering our initial question of whether auditory reduction type has an effect on the pattern of talker differences, no significant interaction was found, indicating that across talker differences cannot be attributed to how the auditory signal is reduced. In hindsight, this is an important finding because it shows that talker differences occur because of fundamental differences between talkers and how they produce speech, and are relatively robust to the different ways in which the signal can be reduced. In addition to answering that question, the present study creates a foundation on which various additional analyses can be applied to further examine talker differences in audiovisual integration. Some of these are described below.

Dietrich (2008) addressed specific information in the auditory signal that impact audio-visual performance by looking at formant transitions in 2-channel filtered speech stimuli for 2 talkers, one much more intelligible than the other. She found substantial differences, particularly in F2 transitions. A similar analysis focusing on formant transitions in each of the 5 talkers in the present study in both the 2-channel and sine wave speech conditions, might be a valuable starting point in examining differences in the talkers' auditory productions.

Results in Tables 1 and 2 show substantially poorer performance for sine wave speech stimuli that have greater high-frequency components (e.g., voiceless stops and fricatives: pat, cat, tat and sat). These confusion matrices suggest that further auditory analyses might focus on stimuli with substantial high frequency components to differentiate sine wave and 2-channel

performance. Differences could be examined in specific perceptual confusions across talkers, by analyzing the individual confusion matrices included in the Appendix. Selected stimuli in these confusion matrices could also be analyzed across talkers acoustically (i.e., formant transitions) to evaluate the acoustic characteristics that facilitate audio-visual integration.

Figures 2 and 3 indicated variability across talkers in audiovisual integration. Possible differences across talkers, in terms of integration, could also be attributed to visual characteristics of the talkers. One question that has not been completely answered is whether a highly intelligible visual stimulus facilitates integration more or whether integration is better facilitated with an ambiguous visual stimulus. Talker DA is interesting in this sense because of the unique way in which he articulates. Anecdotally, listeners consistently complained that DA's speech was difficult to perceive visually because of the odd way he produced syllables. This was true despite the fact that DA's visual intelligibility was not substantially lower than that of the other talkers. Visual characteristics of DA, such as lip opening, lip rounding, and jaw movement, should be evaluated and compared with the characteristics of other talkers to determine if talker differences can be attributed to visual cues of a talker. Specifically, timing of lip separation, extent of mouth opening and face symmetry should be examined.

To gain a more complete understanding of how characteristics of the stimulus facilitate or impede integration, a combination of analyses should be pursued. Specific visual characteristics of talkers may answer questions about the ambiguity of the visual stimulus and how it facilitates integration. Further, the specific confusions across talkers point to specific visual characteristics that should be examined. Acoustic analyses should also be driven by confusion data.

Results of this study have implications for a better understanding of the underlying reasons for talker differences in audiovisual integration, the factor of integration that has received the least amount of attention. In further understanding talker differences, we can better understand the process of how speech is perceived and integrated. With this understanding, we will be able to better understand how to train audiovisual integration, which will help in the design and development of aural rehabilitation programs for individuals with hearing impairments. The present study, consistent with past research, found significant across-talker differences indicating the importance of incorporating multiple talkers into audiovisual integration training. Also, given what we found, the importance of training auditory + visual stimuli is evident in the design of appropriate aural rehabilitation programs.

Chapter 5: References

Anderson, C. (2007). Auditory and visual characteristics of individual talkers in multimodal speech perception. *Senior Honors Thesis, the Ohio State University.*

Andrews, B. (2007). Auditory and visual information facilitating speech integration. *Senior Honors Thesis, the Ohio State University.*

Behrman, A. (2007). Speech and voice science. Plural Publishing.

Braida, L.D. (1991). "Crossmodal integration in the identification of consonant segments," *Q. J. Exp. Psychol.* 43A (3), 647-677.

Dietrich, K. (2008). Analysis of Talker Characteristics in Audio-Visual Speech Integration. *Senior Honors Thesis, the Ohio State University.*

Feleppelle, N. (2008). The role of the auditory signal in auditory-visual integration. *Capstone Project, the Ohio State University.*

Fisher, C.G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 12, 796-804.

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America*, 112 (1), 30-33.

Grant, K.W. & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. The Journal of the Acoustical Society of America, 04 (4), 2438-2449.

Hiss, M. (2008). Audiovisual Integration of Reduced Information Speech Stimuli. *Senior Honors Thesis, the Ohio State University*.

Huffman, C. (2007). The Role of Auditory Information in Audiovisual Speech Integration. *Senior Honors Thesis, the Ohio State University*.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review*, 90 (5), 99-114.

Massaro, D.M. (1987). Speech perception by ear and eye: A paradigm for psychology inquiry. Hillsdale, NJ: Lawrence Erlbaum.

Massaro, D.M. (1998). *Illusions and issues in bimodal speech perception*. in Auditory-Visual Speech Processing Conference. Terrigal, Sydney, Australia. P. 21-26.

Massaro, D.W., and Cohen, M.M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception. *Journal of the Acoustical Society of America*, 108, 784-789.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.

Munhall, K.G., Kroos, C., Jozan, C., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perceptions & Psychophysics*, 66 (4), 574-583.

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-950.

Schow, R.L., & Nerbonne, M.A. (2007). Introduction to audiologic rehabilitation [rev. ed.] Boston, MA: Pearson Education, Inc.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303-304.

Shannon, R.V., Zeng, F.G., Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *The Journal of the Acoustical Society of America*, 104 (4), 2467-2475.

Smith, Z., Oxenham, A., Delgutte, B. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 2002 Mar 7; 416(6876): 87-90.

Tamosiunas, M. (2007). Auditory-Visual Integration of Sinewave Speech. *Senior Honors Thesis, the Ohio State University*.

Tables

Table 1. Confusion matrix averaged across talkers and listeners for the 2-channel condition

Response

Stimulus	2-Channel													
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	bdat	pcat	pdat	bgat
bat	**72.6**	12.9	14.5								0.2			
pat	18.8	**70.3**	8.8		0.6		0.3				0.4	0.1	0.1	
mat	14.5	8.3	**77.5**						0.3					0.1
gat	0.1			**63.8**	16.7		1.5	0.1	1.7	14.7				
cat	0.1	0.6		10.0	**71.7**	0.1	14.8	0.4	0.9	1.1	0.1	0.1		
zat		0.1		1.9	1.2	**60.6**	3.8	14.9	3.1	19.1				
tat	0.3	0.4		2.8	24.1	3.2	**54.3**	6.6	0.6	8.0				
sat				0.5	0.9	5.7	3.1	**88.0**	0.3	1.5				

Table 2. Confusion matrix averaged across talkers and listeners for the sine wave speech condition

Response

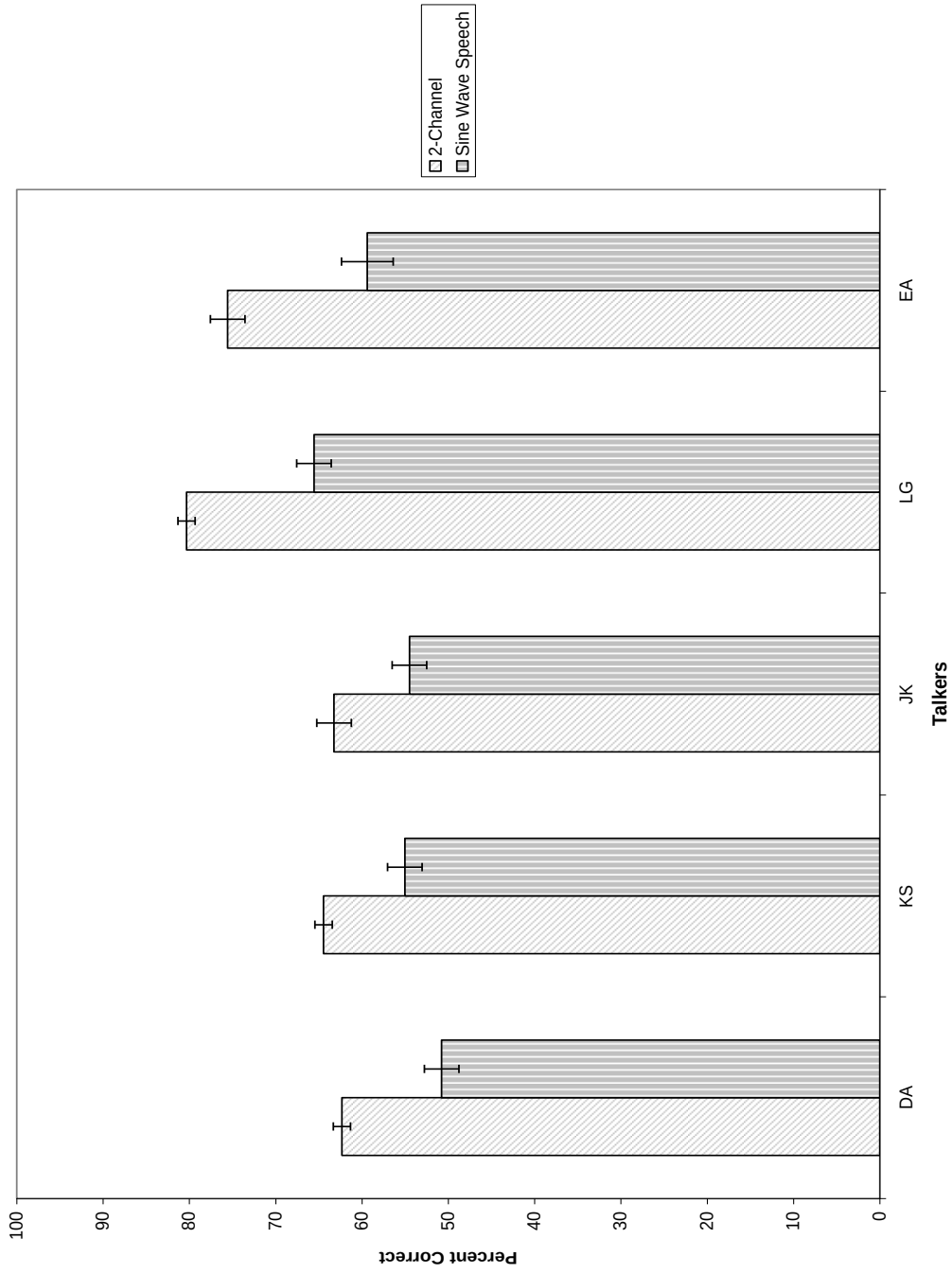
Stimulus	Sine Wave Speech											
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	bdat	pcat
bat	**69.8**	15.4	13.9					0.2		0.2		
pat	27.3	**38.1**	33.8		0.3			0.1			0.1	
mat	9.0	3.7	**87.3**									
gat				**65.6**	17.4	2.9	1.3	1.1	4.9	7.0		
cat		0.3	0.3	34.9	**48.5**	2.9	1.7	1.1	7.7	2.5		0.1
zat			0.1	1.3	0.8	**77.0**	1.5	12.6	1.8	4.7		
tat		0.1	0.4	6.4	17.3	18.9	**14.7**	22.7	9.1	5.1		
sat			0.1	2.2	1.2	32.2	1.9	**59.3**	0.3	3.7		

Table 3. Number of McGurk-like responses out of total number of "other" responses for each talker and each type of speech.

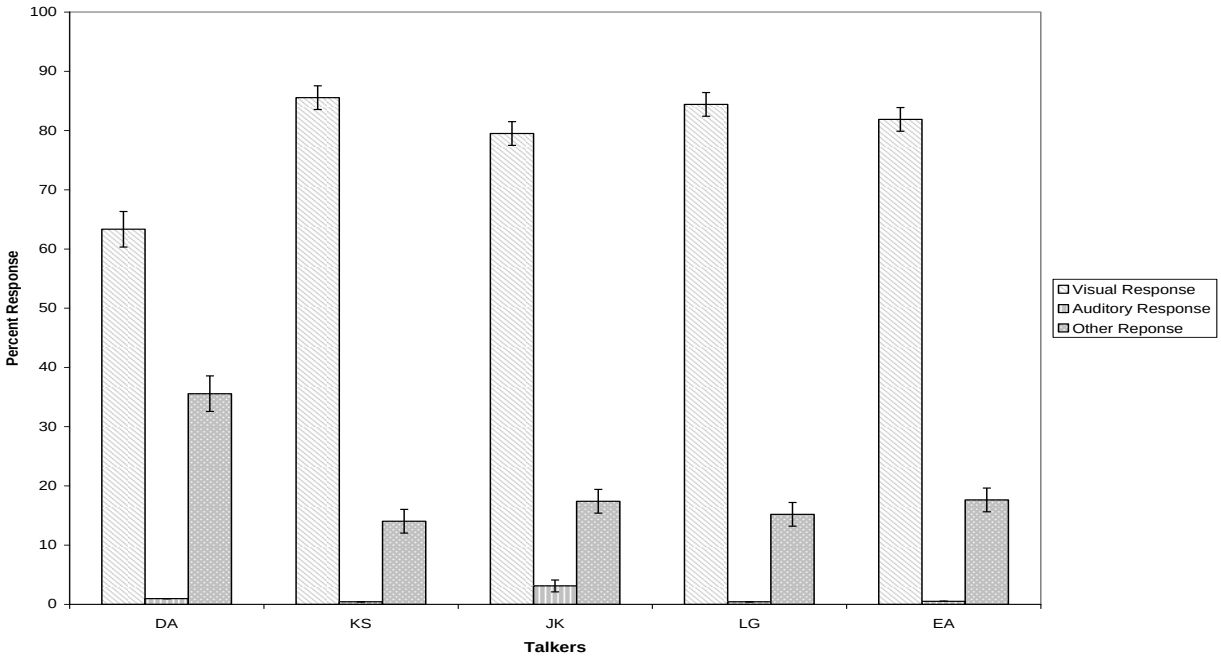
	2-channel filtered speech	Sine wave speech
DA	**386**/403 - 95.78%	**300**/310 - 96.78%
EA	**206**/209 - 98.57%	**376**/380 - 98.95%
JK	**204**/205 - 99.51%	**241**/243 - 99.17%
LG	**189**/190 - 99.47%	**164**/165 - 99.39%
KS	**164**/166 - 98.79%	**128**/132 - 96.97%

List of Figures

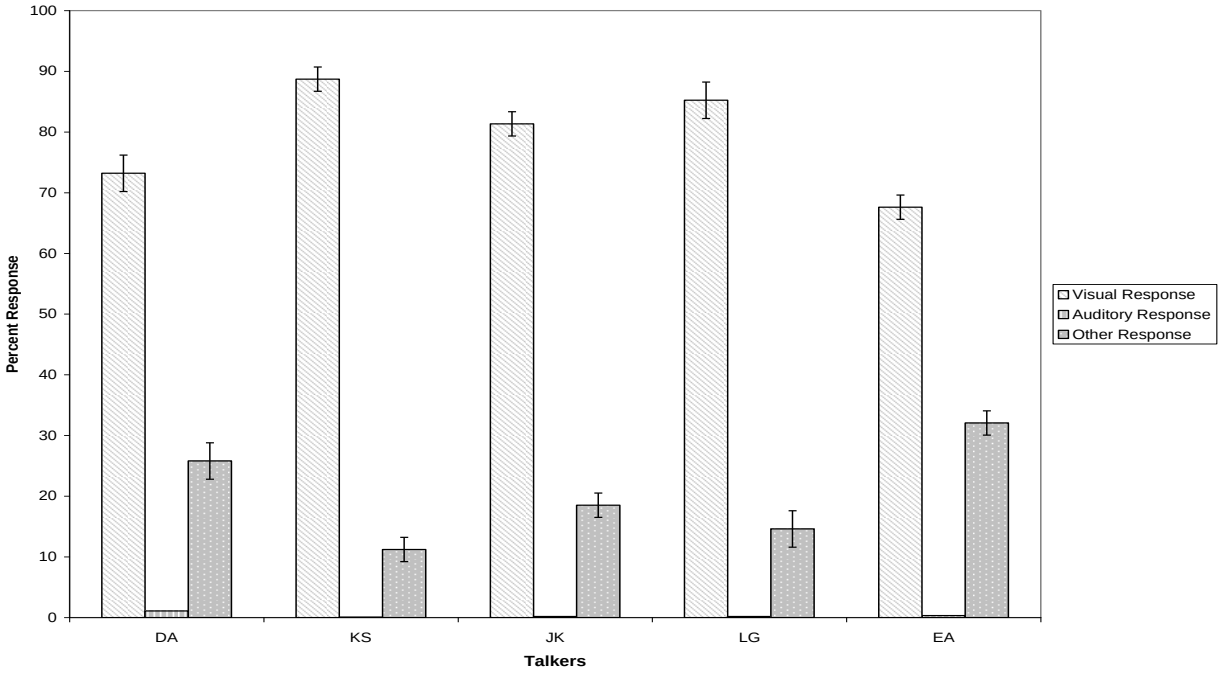
Overall Percent Correct
Figure 1



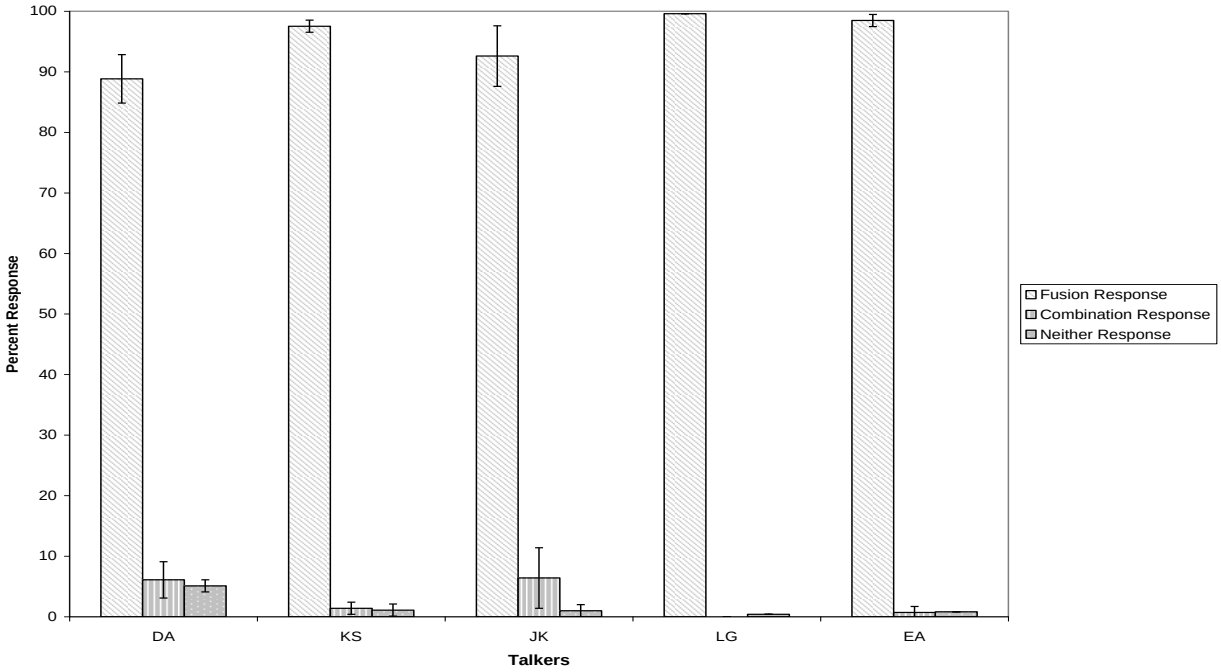
Modality Reliance
2-Channel
Figure 2



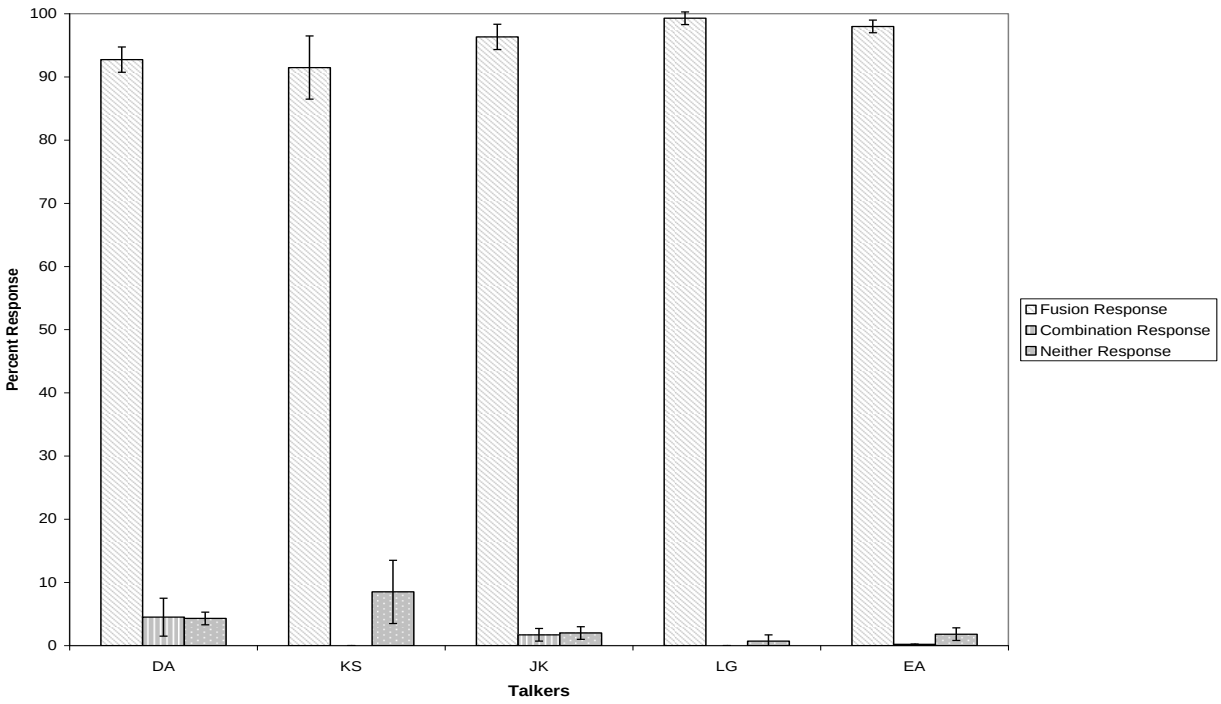
Modality Reliance
Sine Wave Speech
Figure 3



Percent McGurk Response
2-Channel
Figure 4



Percent McGurk Response
Sine Wave Speech
Figure 5



Appendix

List of Appendix Tables

2-channel filtered speech confusion matrices by talker

Table A1: Confusion matrix for Talker DA in 2-channel filtered speech condition

DA 2-Channel												
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	bdat	pdat
bat	**78.7**	8.1	13.2								0.7	
pat	39.0	**30.8**	24.0				0.7				2.1	0.7
mat	25.2	13.6	**64.6**									
gat				**65.6**	4.1		2.7		0.7	26.7		
cat	0.7	1.5		31.9	**50.4**		8.9		0.7	8.9	0.7	
zat						**78.8**		19.0		2.2		
tat	0.7	0.7		5.5	22.6	1.4	**39.0**	1.4	2.7	26.0		
sat						1.4		**98.6**				

Table A2: Confusion matrix for Talker KS in 2-channel filtered speech condition

KS 2-Channel										
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat
bat	**51.7**	25.0	23.3							
pat	23.8	**65.4**	10.8							
mat	18.1	15.0	**66.9**							
gat				**62.5**	26.2		2.5		5.0	3.7
cat				14.2	**82.4**		0.7		2.7	
zat				1.9	0.6	**60.1**	3.8	17.7	0.6	15.8
tat					19.5	5.0	**57.9**	16.4		1.3
sat				0.6	2.5	8.9	13.3	**69.0**	1.3	3.8

Table A3: Confusion matrix for Talker JK in 2-channel filtered speech condition

JK 2-Channel												
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	pcat	bgat
bat	**65.8**	20.8	13.3									
pat	16.7	**77.5**	4.2		0.8						0.8	
mat	20.0	6.2	**73.1**									0.6
gat	0.6			**43.1**	15.0		1.2	0.6	1.2	38.1		
cat		0.7		0.7	**88.6**		8.1		0.7		0.7	
zat		0.6		5.7	3.8	**40.3**	10.1	0.6	8.2	30.8		
tat				6.9	46.2		**33.1**	1.2	0.6	11.9		
sat				0.6	0.6	3.1	1.2	**91.9**		2.5		

Table A4: Confusion matrix for Talker LG in 2-channel filtered speech condition

LG 2-Channel										
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat
bat	**84.6**	3.1	13.1							
pat	2.0	**97.3**			0.7					
mat	3.1	3.1	**93.7**							
gat				**66.7**	33.3		0.7		1.3	1.3
cat					**83.8**	0.8	15.4			
zat						**61.9**	2.5	18.7	4.4	12.5
tat	0.6	0.6			22.6		**78.0**	0.6		
sat					1.2	9.4	0.6	**88.1**		1.2

Table A5: Confusion matrix for Talker EA in 2-channel filtered speech condition

EA 2-Channel												
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	bdat	pdat
bat	**80.0**	9.2	10.0									
pat	13.3	**80.0**	4.7		1.3		0.7					
mat	6.2	3.7	**88.7**						1.2			
gat				**83.6**	2.9					2.1		
cat		0.7		3.6	**51.4**		42.1	2.1				
zat				1.9	1.3	**64.2**	1.9	18.9	1.9	10.1		
tat		0.6		1.9	9.4	9.4	**62.5**	13.1		2.5		
sat				1.2		5.6		**93.1**				

Sine wave speech confusion matrices by talker

Table A6: Confusion matrix for Talker DA in sine wave speech condition

DA SWS										
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat
bat	**53.5**	22.8	20.0	2.1		0.7				0.7
pat	30.6	**45.3**	21.3		0.7			0.7		
mat	30.2		**67.8**							
gat				**42.7**	17.4	6.5	2.2	2.9	5.8	22.5
cat		0.7		33.3	**45.6**	6.8	2.7	2.0	2.0	5.4
zat				0.7		**94.6**		2.0	1.3	1.3
tat		0.7		9.6	48.9	12.6	**7.4**	10.4	0.7	9.6
sat			0.7	2.1	2.1	36.9	2.1	**53.2**	0.7	5.7

Table A7: Confusion matrix for Talker KS in sine wave speech condition

KS SWS										
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat
bat	**60.0**	19.2	20.8							
pat	36.6	**39.2**	26.9							
mat	4.4	12.5	**83.1**							
gat				**68.2**	21.0	1.9	1.3	0.6	5.1	1.9
cat			0.7	39.4	**51.8**	1.5	0.7	0.7	6.6	0.7
zat				0.6	2.5	**63.1**	1.9	29.4	0.6	1.9
tat			0.6	1.3	6.3	19.6	**17.1**	49.4	3.2	2.5
sat				5.6	1.2	32.5	2.5	**56.9**		1.2

Table A8: Confusion matrix for Talker JK in sine wave speech condition

JK SWS											
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	bdat
bat	**79.2**	15.4	4.6					0.8			
pat	16.9	**43.1**	39.2								0.8
mat	3.7	1.9	**94.9**								
gat				**73.6**	20.3	2.0	0.7		2.7	0.7	
cat				37.8	**43.2**	4.7	2.0	0.7	11.4	0.7	
zat			0.6	1.9	1.3	**59.5**	5.7	20.3	5.7	5.1	
tat			0.6	12.7	14.6	15.2	**15.2**	17.7	19.0	5.1	
sat				1.3		53.2	3.2	**32.3**		10.1	

Table A9: Confusion matrix for Talker LG in sine wave speech condition

LG SWS										
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat
bat	**80.8**	11.7	7.5							
pat	24.7	**38.7**	36.7							
mat	3.7	3.7	**92.5**							
gat				**68.4**	24.3		1.5		5.9	
cat				34.0	**58.5**	0.7	2.0		2.0	2.7
zat				3.2		**84.1**		7.0	0.6	5.1
tat			0.6	2.5	15.1	30.2	**18.2**	2.5	3.8	3.8
sat						12.5	0.6	**85.6**		0.6

Table A10: Confusion matrix for Talker EA in sine wave speech condition

EA SWS											
	bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	pcat
bat	**76.2**	7.9	15.8								
pat	29.5	**23.7**	45.3	1.4	0.7						
mat	4.4		**95.6**								
gat				**74.1**	3.6	4.3	0.7	2.2	5.6	10.8	
cat		0.8	0.8	30.1	**43.6**	0.8	0.8	2.3	17.3	3.0	0.8
zat						**84.9**		3.8	0.6	10.1	
tat				6.6	6.0	15.9	**14.6**	32.5	17.9	5.3	
sat				1.9	2.5	26.8	1.3	**67.5**	0.6	1.3	